Vol. 4, No. 2, Fall 1974

SPEECH CHAIN AS AN ANALYSIS-BY-SYNTHESIS MODEL: A REVIEW

Farid M. Onn

0.0. Introduction

There are two aspects of ordinary speech communication: a linguistic message is encoded into articulatory gesture by a speaker, and acoustical signals are interpreted as a linguistic message by a hearer. Different messages are encoded as different gestures, and different signals are interpreted as different messages. Exactly how articulatory gestures are related to acoustical (let alone neurophysiological) signals, and what actually happens in the speech mechanisms during the production of speech events, have been some of the primary concerns of speech scientists for many years.

The notion that there is a variation of the gesture or the signal in accordance with the linguistic code of a particular language is based on the fact that there is no a one-to-one relation[1] between code and gesture, signal and message, or between segments of an utterance and the phonemes. It is also because of this lack of a one-to-one correspondence between segments and phonemes that many phoneticians have expressed the view that speech analysis procedure which postulates that the hearer first segments the utterance and then identifies the individual segments with particular phonemes, can not successfully be implemented.[2] One could foresee this failure in the light of the complexity in which perceived language is related to the acoustical signal which conveys it; also, the changing configurations of man's vocal-tract, which are specified in terms of phonetic parameters, have been observed to be the result of instructions not from a single phoneme but from a given sequence of phonemes. However, to establish and study this complex relation between perceived language and the acoustical signal is by no means a hopeless undertaking. In fact, another speech analysis procedure has been proposed, a model that has been described as having the power to transform the continuously-changing speech signal into a discrete output without relying

crucially on segmentation. This model which analyzes the internally generated speech patterns through active internal synthesis of comparison of signals, has been called analysis-by-synthesis.

This brief paper attempts to characterize speech chain in terms of the proposed analysis-by-synthesis model, and to examine some of the empirical descriptive problems that a speech processing model, such as this, will encounter.

1.0. Traditional view

Traditionally, a model of speech production is understood in terms of the view that the direct input of the speech production system consists of a series of phoneme commands. In expressing such a view, Halle (1962) writes:

> It is assumed that stored in the memory of the speaker there is a table of all the phonemes and their different actualizations. This table is basically a dictionary in which can be found the different vocal-tract configurations or gestures that are associated with each phoneme, and the conditions under which each of the configurations or gestures is to be used. Associated with some phonemes there may be but a single configuration or gesture; with others the number of gestures may be large...In producing an utterance the speaker looks up in the table each phoneme in the utterance and then causes his vocal-tract to assume in succession the configurations or gestures corresponding to the phonemes composing the utterance. The vocal-tract behavior in turn causes disturbances in the air which are transmitted to our ears as acoustical signals (p. 429).

In short, the traditional view of the process of speech production assumes that there is in the speaker a set of instructions /rules which permit him to transform a sequence of discrete entities of phonemes into quasi-continuous behavior of the vocal-tract and later into a quasi-continuous acoustical signal. However, a speech production model which postulates that the hearer first segments the utterance and then identifies the segments as particular phonemes requires that he possesses in his memory

a list of the acoustical equivalents of the phonemes and that he must also be able to segment all utterances. But in principle, as Halle has observed, given the acoustic input, it is not possible for a hearer to segment all utterances. Analysis-by-synthesis, as a speech processing model, has been claimed to be a more effective model which could, among other things, overcome the problems which result from the seemingly impossible task of achieving complete segmentation of all utterances.

2.0. Analysis-by-synthesis model

Analysis-by-synthesis involves a process of specifying an unknown sign in terms of a best match selection from a standard inventory. More specifically, the model postulates that the process of speech chain involves the internal synthesis of patterns according to certain rules and a matching of these patterns which are internally generated against the pattern under analysis.

The block diagrams shown in figures I and II illustrate the basic operations involved in the speech analysis procedure to be described. The diagrams are essentially those of Stevens (1960) and Halle and Stevens (1964), except that they have slightly been simplified. The operation of stage I and stage II completes the proposed model of speech chain.

In figure I, the input speech signal first undergoes a "preliminary analysis" which constitutes various transformations, such as segmentation, identification of segments by special attributes, etc. Following this "preliminary analysis", the signal is then sent to the "comparator", where it is compared with the stored signals. The "comparator" also establishes a measure of the error between the stored articulatory descriptions and those generated by the model. Any such error is readily channeled through the "control" element which, after a number of trials, generates an output that identifies the acoustical signal. The "rules"[3] which may be regarded as the core of the speech process model transforms phoneme sequences to phonetic parameters. These rules also operate on the input signal to yield instructions to the vocal mechanism, and these instructions cause appropriate activity in the articulatory mechanism to generate the output sound

STAGE 1:

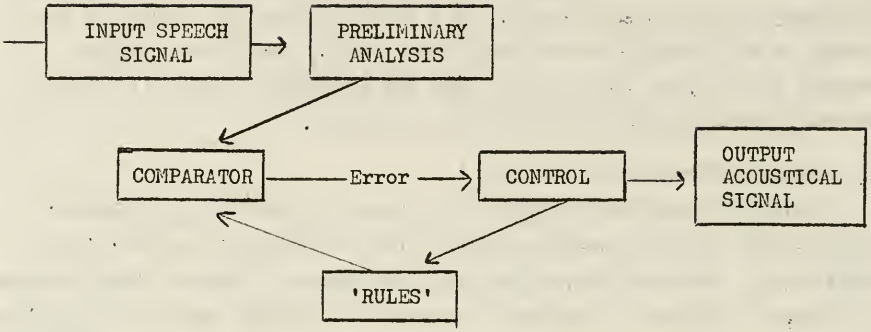

Fig. 1. Model of Speech Production

STAGE 11:

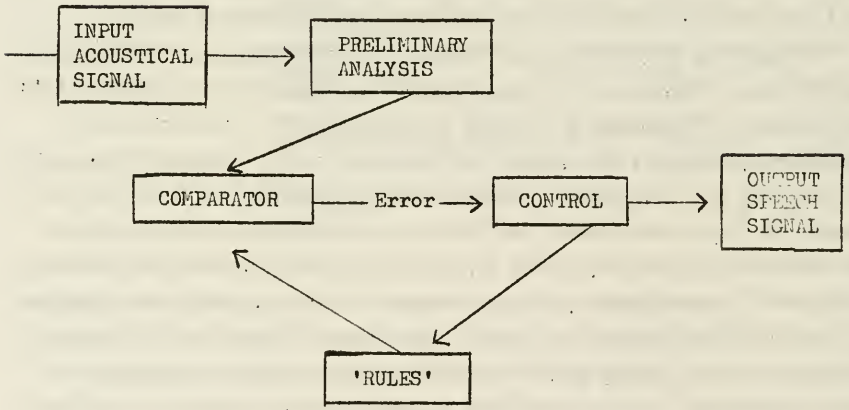

Fig. 11. Model of Speech Perception

which later becomes the input acoustical signal.

The operations carried out in the second stage of the speech process, as shown in Fig. II, are a repetition of those in the first stage, except that in the second stage, the input is an acoustical signal. Like the rules operating in the first stage, those in the second stage also constitute the heart of the perception model. In fact, following Halle (1962), the dual processes of production and perception ought to be viewed as separate utilization of a common core of rules rather than as distinct processes each with its own body of rules.[4] The two separate stages of the speech analysis model may be combined to form a truly single process of speech chain, as shown in Fig. III. In such a model, the group of components performing the functions of "preliminary analysis", "comparison" and "control" in a single block, has been labeled the "strategy".[5]

It must be emphasized that it is the employment of the rules, which are recursive, in the model that makes analysis-by-synthesis more effective and powerful than other previously constructed speech analysis procedures which compile speech generally from a dictionary of recordings or a look-up table of values. Obviously the operation of the rules of systematic synthesis is much more complex than a table of look-up procedure. As Kim (1966) describes it, "this set of rules may be regarded as a computer program instructing the synthesizer what parameters to operate, to what degree, for how long, etc., when given phonetic categories." (p. 63). The importance of "rules of synthesis" has also been greatly emphasized by Stevens (1960) who says that "rules for generating spectral patterns rather than the entire catalog of patterns themselves are stored, with a resulting large saving in storage capacity. Furthermore, if a proper strategy is devised for selecting the order in which patterns are synthesized for comparison with the input, then the number of patterns which must be generated and compared may be of orders of magnitude less than the total number of patterns that could be generated by the rules." (p. 53). What this amounts to is that not only the rules here are important but the order in which patterns are synthesized is also crucial, for without such order, the idea of analysis-by-synthesis

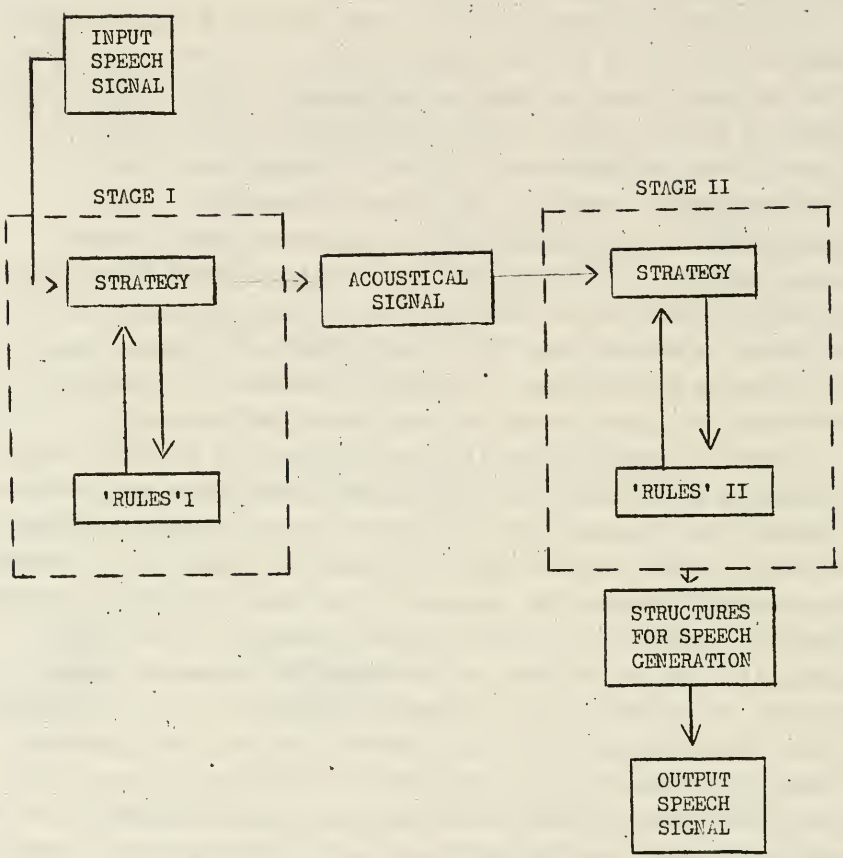

Fig. III Speech Chain as Analysis-by-Synthesis Model

becomes merely a process of trial and error. In proposing the kind of strategy for selecting the order, Stevens (1960) says, "the order in which different articulatory descriptions are tried may depend in part on data from a preliminary analysis of the signal, in part on data from previous spectra, and part on the results of previous trials on the spectrum under analysis." (p. 50).

3.0. Problems of selection

We have attempted to show that the generative mechanism operating in an analysis-by-synthesis speech model, which may need a digital computer of reasonable size and speed to stimulate all the operations, is more effective than a model which stores the pre-recorded fixed values in a look-up table, and drawing these values from the table by a simple substitution procedure. However, despite the complexity of the analysis-by-synthesis model, and its important advantages over other speech analysis procedures, the model is still a long way from being a complete machine capable of describing the whole of human speech behavior. There are still unsolved problems, mainly those that arise from our lack of knowledge in some areas of the speech behavior. For example, there is the problem of selection of the detailed form of representation of the articulatory description, especially at the neurophysiological level. Also, there is the requirement that in order for the generative rule components in the model to produce their maximum functions, they must have a complete set of rules capable of generating, for example, an articulatory description from a sequence of phonetic symbols; and also rules that describe the conversion of phonetic parameters to time-varying speech spectra. The following series of relations involving such generative rules have been noted: those between (1) the phonetic parameters and the vocal-tract geometry and excitation characteristics, (2) the transformation from vocal-tract geometry to the transfer function in terms of poles and zeros, and (3) the conversion from the pole-zero configurations and pertinent excitation characteristics to the speech spectra.

The problems that may be associated with the second set of rules are not without complexity. There is, for example, the problem of the utilization

of those phonetic parameters that are not governed by the language in question which must be described by the rules. The rules must also be able to specify the transformation from phoneme level representation having a discrete characteristic to continuous signals resulting from the inertia of the neural and muscular structures involved in speech production. In their discussion on the model and the notion of distinctive features, Stevens and Halle (1967) made particular reference to the question of the difficulty to isolate the segments and features in the actual speech event. They believe, however, that this difficulty could be resolved by recognizing explicitly that "...characterizations of speech in terms of segments and features are not more or less naturalistic records of particular physical events but are rather abstract representations of classes of events." (p. 90). In other words, it may be argued that an abstract representation of the speech event and a set of appropriate generative rules are involved in the process of speech production, and that those segments and features of the abstract representation may be regarded as instructions for particular types of behavior of the speech-generating mechanism. When these instructions are carried out, the various reactions occurring between different physiological structures will yield a quasi-continuous gesture in which the discrete instructions initiating the gesture are no longer always observable as distinct components. Finally, the execution of these instructions produces the acoustical signal.

It is therefore evident that the speech-generating mechanism can also be explained with reference to the nature of the abstract framework of segments and features, the entities which underlie the whole phonology of every human language. I believe it is also this portion of generative phonology, i.e., analysis of segments and features, that the first rule component in the analysis-by-synthesis model is to be mainly associated with. In particular, it is the conversion of the abstract representation of segments and features into a sequence of phonetic symbols that constitutes the major function of the first rule component in the speech analysis model, whereas the conversion of these phonetic symbols into phoneme level representation and later into words and sentences, in that order, is mainly the function of the second rule

component. In short, the first set of rules deals with the lower level of speech description, while the second set involves the higher level of analysis.

4.0. Hierarchy and well-formedness

Perhaps, it is pertinent at this point to touch upon the question of "hierarchy" and "well-formedness". It has often been asserted that language is composed of segments and that these segments are arranged in hierarchically ordered layers.[7] Also it is a well known fact that every language possesses some specific constraints on the sequence of segments that can constitute a well-formed utterance. Language structure is said to be hierarchically ordered if we look at it as constituting two grammars: the phonology which contains segments that are themselves empty of meaning; and the other comprising morphology and syntax, which ascribes meaning and structural well-formedness to phonological segments. The proposed analysis-by-synthesis speech model can be said to have been developed to meet the presence of such linguistic phenomena. Thus, the notion of hierarchial order in language structure, for example, has necessitated the development of two rule-components in the proposed speech analysis model. So that, while the rules in the first component can take care of the lower order of phonological representation, the rules in the second component will, among other things, specify those constraints on the sequence of segments that constitutes well-formedness.

5.0. Conclusion

In the preceding paragraphs, we have attempted to characterize speech chain in terms of the proposed analysis-by-synthesis model. Some of the empirical descriptive problems involving such a speech processing model have been examined. The problems are mainly those which involve or are associated with the rules that operate in the model. The model, however, is considered to have important advantages over other speech processing techniques. It has also been suggested that a model of the type reviewed here has applications in the analysis of linguistic phenomena at various levels of representation: acoustic, phonological, morphological, and syntactic.

FOOTNOTES

[1]For discussion of some of the evidence, see P. Ladefoged's contribution to the Teddington Symposium, The Mechanization of Thought Processes, National Physical Laboratories, Symposium #10 (London 1959). C.G.M. Fant (1962) has also observed that the concept of speech is not as a sequence of discrete units with distinct boundaries, but rather as a continuous succession of gradually varying and overlapping patterns. He says, "the number of successive sound segments within an utterance is greater than the number of phonemes...Sound segment boundaries should not be confused with phoneme boundaries. Several adjacent sounds of connected speech may carry information on one and the same phoneme, and there is overlapping in so far as one and the same sound segment carries information on several adjacent phonemes." (p. 9).

[2]In his discussion on the acoustic aspects of speech, C.G.M. Fant (1962) commented on the failure of such speech analysis procedure, with the following words: "Phoneme recognizing machines of a simpler analog type have been constructed but their performance has not been very advanced. The possible vocabulary or phoneme inventory has been restricted, and the machines have not responded very well to any one else than "his master's voice"." (p.3).

[3]In Halle (1962), and Halle and Stevens (1964), these "rules" are referred to as "generative rules". But Stevens in his earlier paper (1960), refers to them as simply the "rules". Similarly, Liberman et al. (1959) made reference to "rules of synthesis" by saying: "The place rule for /l/ specifies locus frequencies at 360, 1260, and 2880 cps. . . the place rule for /æ/ fixes formant frequencies at 750, 1650, and 2460 cps.... (p. 1497).

[4]See M. Halle (1962), p. 433

[5]See M. Halle and K.N. Stevens (1964) p. 610.

[6]Ibid., p. 611.

[7]This view has been expressed by linguists like Liberman, Cooper, MacNeilage, and Kennedy (though one might add that not all linguists will buy the hierarchical part of their assertion). See Liberman et al. (1967) p. 69.

REFERENCES

Fant, C.G.M. 1960. Acoustic Theory of Speech Production, Mouton, The Hague.

_____. 1962."Descriptive Analysis of the Acoustic Aspects of Speech", LOGOS, Vol. 5, No. 1, pp. 3-17.

_____. 1968. "Analysis and Synthesis of Speech Processes", in Manual of Phonetics, 2nd. ed. by B. Malmberg (ed.), North-Holland Pub. Co., Amsterdam, pp. 173-277.

Halle, M. 1962. "Speech sounds and sequences", in Proceedings of the Fourth International Congress of Phonetic Sciences, Mouton, The Hague, pp. 428-434.

_____ and K.N. Stevens. 1964. "Speech recognition: a model and a program for research", in The Structure of Language: Readings in the Philosophy of Language, J.A. Fodor and J.J. Katz (eds.), Prentice-Hall, Englewood Cliffs, pp. 604-612.

Kim, C-W. 1966. "The role of a speech synthesizer", in UCLA Working Papers in Phonetics, No. 5, pp. 16-26.

Liberman, A.M., F.S. Cooper, Katherine S. Harris, P.F. MacNeilage, and Studderd Kennedy. 1967. "Some observations on a model for speech perception", in Models for the Perception of Speech and Visual form, W. Wathen-Dunn (ed.), The MIT Press, Cambridge, pp. 68-87.

Stevens, K.N. 1960. "Toward a model for speech recognition", JASA 32, pp. 45-55.

_____ and M. Halle. 1967. "Remarks on analysis by synthesis and distinctive features", in Models for the Perception of Speech and Visual form, W.Wathen-Dunn (ed.), The MIT Press, Cambridge, pp. 88-102.